Review

Javed Khan
Lao H. Saal
Michael L. Bittner
Yidong Chen
Jeffrey M. Trent
Paul S. Meltzer

Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Expression profiling in cancer using cDNA microarrays

Currently there are over 1 000 000 human expressed sequence tag (EST) sequences available on the public database, representing perhaps 50–90% of all human genes. The cDNA microarray technique is a recently developed tool that exploits this wealth of information for the analysis of gene expression. In this method, DNA probes representing cDNA clones are arrayed onto a glass slide and interrogated with fluorescently labeled cDNA targets. The power of the technology is the ability to perform a genome-wide expression profile of thousands of genes in one experiment. In our review we describe the principles of the microarray technology as applied to cancer research, summarize the literature on its use so far, and speculate on the future application of this powerful technique.

Keywords: cDNA microarray / Gene expression / Cancer research / Review EL 3273

Contents

1 Introduction

The structure and biological behavior of a cell is determined by the pattern of gene expression within that cell. Each human cell contains approximately three billion base pairs, which encode between 50 000 to 100 000 genes [1–3]. In any given cell only a small fraction of these genes is being actively transcribed. Cancer can be regarded as a genetic disease occurring as a result of progressive accumulation of genetic aberrations [4]. Neoplastic cells have numerous acquired genetic abnormalities including aneuploidy, chromosomal rearrangements, amplifications, deletions, gene rearrangements, and loss or gain of function mutations. These changes result in a deviation of the gene expression profile from that of the normal progenitor cell. This perturbation leads to the aberrant behavior common to all cancer cells: dysregulated growth, lack of contact inhibition, genomic instability, and propensity for metastasis. With few exceptions, cancer remains an incurable disease, and an increased understanding of the molecular basis of cancer will allow the development of new treatment strategies that will impact positively on prognosis.

The standard techniques of molecular biology have been successfully used to identify increasing numbers of genes involved in cancer. However, these methods are highly focused, targeting only one specific gene or chromosome region at a time, and do not provide insight into global gene expression. With the development of the expressed sequence tag (EST) database there has been a recent shift from “structural genomics” towards “functional genomics”, with genome-wide expression analysis in the forefront of this transition. Three new techniques have emerged in the literature for genome-wide expression analysis, including serial analysis of gene expression (SAGE), DNA microarrays, and oligonucleotide chips.

In the first paper describing the elegant method of SAGE [5], the authors isolated short diagnostic sequence tags from pancreas, which were concatenated, and cloned. They found that subsequent manual sequencing of 1000 tags revealed a gene expression pattern characteristic of pancreatic function. New pancreatic transcripts corresponding to novel tags were also identified. This technique has been further used to identify genes induced by the *p53* tumor suppressor gene [6], and it may be possible to identify ESTs that are differentially expressed in human cancer to generate an expression profile for that

Correspondence: Paul S. Meltzer, MD, PhD, NHGRI/NIH, Building 49, Room 4A10, 49 Convent Drive, Bethesda, MD 20892-4470, USA
E-mail: pmeltzer@nhgri.nih.gov
Fax: +301-402-2040

Abbreviation: EST, expressed sequence tag

 0173-0835/99/0202-0223 $17.50+.50/0

cancer. Although SAGE has the potential to generate genome-scale expression profiles, human cDNA microarrays have the particular advantage that they are readily amenable to the analysis of multiple samples, thereby generating a large amount of gene expression data for statistical analysis.

Gene expression monitoring using microarrays was first described using radioactive targets hybridized onto filter-immobilized cDNA clones. Large-scale cDNA microarrays were used by Drmanac *et al.* [7, 8], who have produced DNA microarrays containing up to 31 104 cDNA clones which were PCR-amplified and robotically spotted onto nylon membranes for gene expression and discovery experiments. DNA microarrays printed on glass and hybridized with fluorescently labeled cDNA are a significant improvement on the filter-immobilized DNA arrays. The technology was first described by Schena *et al.* [9], who printed 48 genes of *Arabidopsis thaliana* onto glass slides and measured differential expression of genes between two different tissues, root and leaf. Fluorescent targets were made from each of these tissues by reverse transcription of mRNA using distinct fluorochromes. By measuring the intensity ratio for each printed gene they were able to show widespread differences in gene expression between these two tissues. The two-color fluorescence detection scheme has the advantage over radioactively labeled targets of allowing rapid and simultaneous differential expression analysis of independent biological samples. In addition, the use of ratio measurements compensates for probe-to-probe variations of intensity due to DNA concentrations and hybridization efficiencies.

The Affymetrix GeneChip^{TM} is produced by using a modification of semiconductor photolithography to synthesize tens of thousands of oligonucleotides onto silicone chips. Using this method it is possible to produce arrays containing more than 65 000 different 20mer oligonucleotides in an area of 1.6 cm^2. Although originally developed for mutation detection, Lockhart *et al.* [10] adapted the same technology to measure expression levels of cytokine genes in murine T cells. They found it a sensitive technique able to measure mRNA levels at a frequency of 1:300 000.

Immobilized oligonucleotide arrays may be regarded as an alternative technology to cDNA arrays. Although the broad principles of both methods are similar, the process of printing cDNA microarrays has the significant advantage of technical feasibility for laboratories engaged in genome research. In this paper we will limit our comments to the cDNA microarray technology developed at the Cancer Genetics Branch of the National Human Genome Research Institute, NIH.

2 Principles of microarray method

2.1 Probe choice and production

Currently, expression arrays containing up to 8000 genes or probes are printed onto a 2 cm × 4 cm area on a microscope glass slide with a probe diameter size of 75–100 μm and 150 μm spacing between probes. With improvements in printing and scanning technology the ultimate goal is to develop arrays which contain the entire human genome. Because of the redundancy in the EST database, efforts have been directed at grouping these sequences into clusters, known as the UniGene set [1, 11]. The entire UniGene collection (http://www.ncbi.nlm.nih.gov/) contains more than 48 000 3′-anchored clusters of sequences, each cluster representing the transcription product of a distinct human gene. We are using a subset of the UniGene set, comprising 15 289 genes (the 15K set), to produce our cDNA microarrays. All sequences selected for inclusion in the 15K set meet at least one of the following criteria: (i) correspondence to a named or functionally cloned gene (6842); (ii) inclusion on the human transcript map (7679); (iii) significant similarity to known proteins in the SwissProt database (1985); (iv) specific research interest of collaborators (365). At least one sequence in each cluster corresponds to a physical cDNA clone that is available from the IMAGE consortium and has been re-arrayed from the original libraries.

2.2 Printing

The cDNA clones are grown in 96-well format, the plasmid DNA is extracted, and the cDNA insert is PCR-amplified using vector primers. The products are ethanol-precipitated, resuspended in 3 × saline sodium citrate (SSC) and printed onto immobilized slides using a custom-built robot. The microscope slides are precoated with poly-L-lysine to enhance DNA binding, as described by Shalon *et al.* [12, 13]. Figure 1 shows a schematic diagram of our arrayer in which 8 “quill” pins, mounted to a cantilever arm, pick up the DNA from each of the 96 wells and print it onto each of the microscope slides in exact predefined positions. Once the DNA is deposited, the slides are washed, preblocked to prevent nonspecific binding of target, denatured, and UV-crosslinked. The slides are then ready for hybridization.

2.3 Target production and hybridization

Total RNA extracted from test and reference cells is fluorescently labeled using oligo dT-primed reverse transcription (Fig. 2) by utilizing nucleotides tagged with either Cy3 or Cy5. The unincorporated fluor-dUTPs are re-

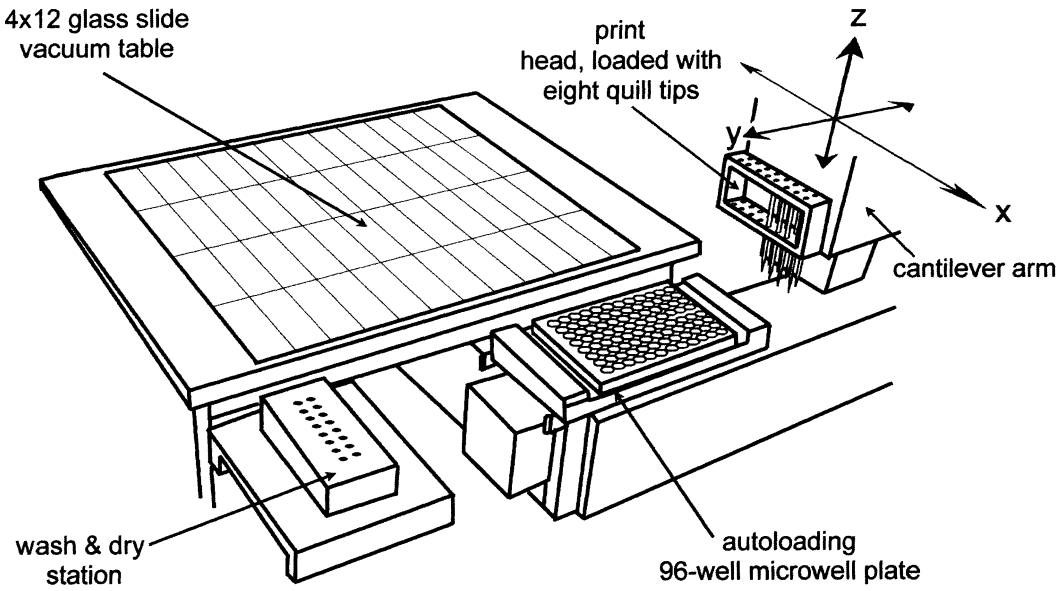

Figure 1. cDNA microarray probe printing apparatus. Computer-controlled robotic cantilever arm, capable of moving in *XYZ* directional planes, can be armed with up to 16 (two rows of eight) "quill" print tips on the print head. In one automated print cycle, the print head dips the quills into a set of probe DNA wells arrayed in 96-well microwell plates; then the print head traverses the vacuum table and touches the quill tips to each glass slide in succession, depositing probe DNA; the print head continues to the wash/dry station where the tips are cleaned twice with water and dried. This cycle repeats as the print head returns to wet the tips in the next set of probes, continuing until all probes of a 96-well microwell plate have been printed. An autoloading mechanism removes spent microwell plates and can serve up new plates. By this method, microarray slides can be printed with as many as 15 000 precise and discrete cDNA probes.

moved, the Cy3 and Cy5 probes combined, and then mixed with blockers consisting of poly (dA), tRNA and Cot1 DNA. The target mixture is hybridized to the probes on the glass slides for 16–24 h at 65 °C, washed, and scanned.

2.4 Detection

Fluorescence intensities at the immobilized probes are measured using a custom-designed laser confocal microscope with a scanning stage (70–90 cm/s) and a photomultiplier tube (PMT) detector. Intensity data are integrated over 15–20 micron square pixels and recorded at 16 bits. The two fluorescent images (Cy3 and Cy5) obtained with the appropriate excitation and emission filters (Fig.3) constitute the raw data from which differential gene expression ratio values are calculated.

2.5 Image analysis and normalization

The two image files generated by the scanner are analyzed using software (DeArray) developed by Chen *et al.* [14]. As each probe is robotically printed to a predefined position, the scanned images can be overlaid with a grid that divides the images into segments, each containing a probe spot. All clone information, including gene name, clone identifier, and source microplate position, is attached to each segment by this process. Each of the images is arbitrarily assigned a pseudocolor (*i.e.*, red for Cy5 and green for Cy3). The target is identified within each segment and the target fluorescent intensity is calculated for each color by averaging the intensities of every pixel inside the detected probe region. The local background intensity of each color is also measured for each spot within each segment. For every spot in each color channel, the final target intensity values are derived by subtracting the local background intensity from the average fluorescent intensity.

Next, a normalization process is performed to compensate for differential efficiencies of labeling and detection of Cy3 and Cy5. The process involves calculating the average intensity, in both color channels, for a set of internal controls consisting of 88 housekeeping genes. These genes are preselected and have been verified on numerous hybridizations as being stable for most experiments (red/green ratio = 1.0). Figure 4 shows the stability of this ratio over a range of mean intensity values for

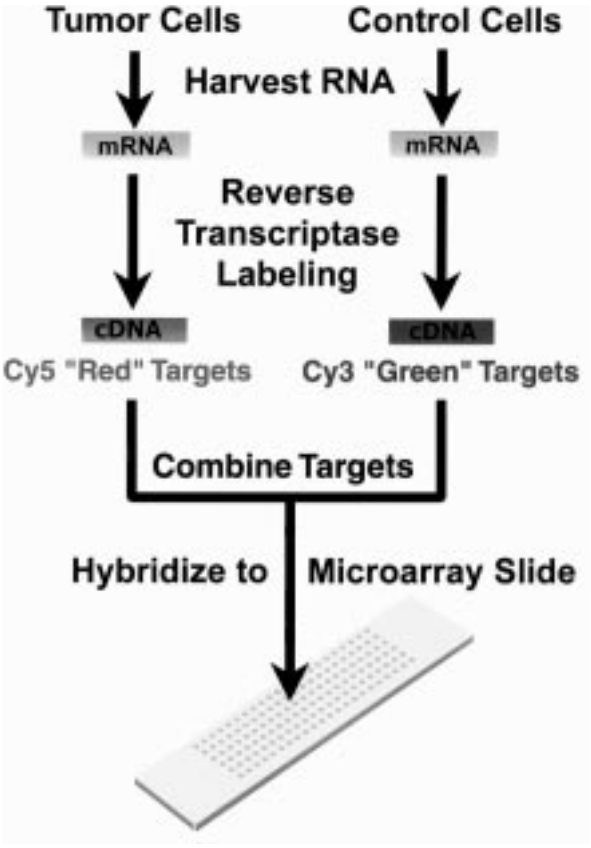

Figure 2. Schematic of target preparation and hybridization

these 88 genes. A normalization constant is then derived and used to calculate a calibrated red/green ratio for each cDNA spot within the image [14]. In addition, the ratio variance of the 88 control genes is used to calculate 99% confidence intervals in which the ratios are considered to be no different from 1. The output of the analysis is in the form of a pseudocolored image of the entire array. Individual spots can be highlighted using the mouse cursor, and information including gene name, clone identity, intensity values, intensity ratios, normalization constant, and user-defined confidence intervals can be obtained. A spreadsheet of the confidence interval outliers containing this information is also generated. Figure 5 shows a microarray experiment after DeArray image analysis. In this example, a normal myofibroblast cell line (labeled green) was compared with a rhabdomyosarcoma cell line (labeled red), and hybridized onto a cDNA microarray containing 1238 elements. Several differences between these two cell lines are observed. All data from each experiment can be downloaded into the ArrayDB [15] database *via* the Internet (http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/), which provides additional tools for comparing data across experiments as well as for extracting data from individual array hybridizations.

3 Sensitivity and specificity

It is estimated that the sensitivity of this method allows the detection of mRNA species comprising 1:10 000 of the mass of poly(A)$^+$ [13]. Comparisons between the micro-

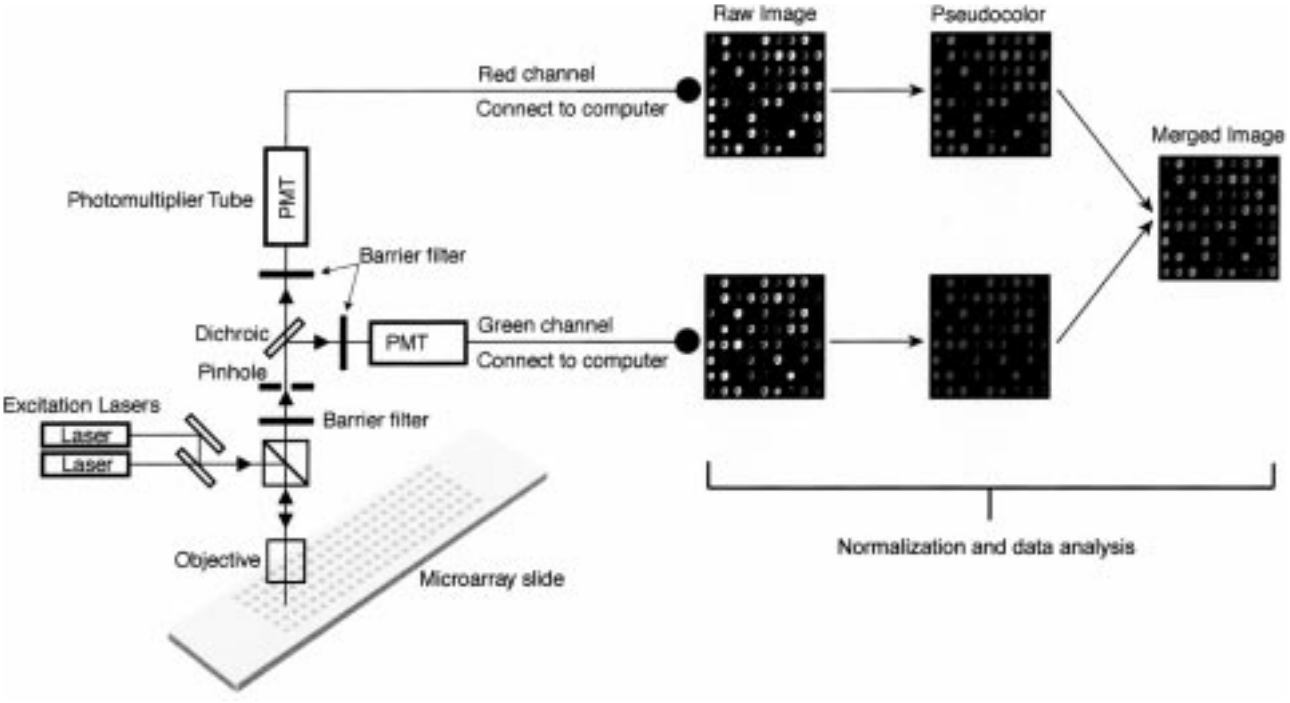

Figure 3. Diagram of confocal laser microscope scanner and image analysis

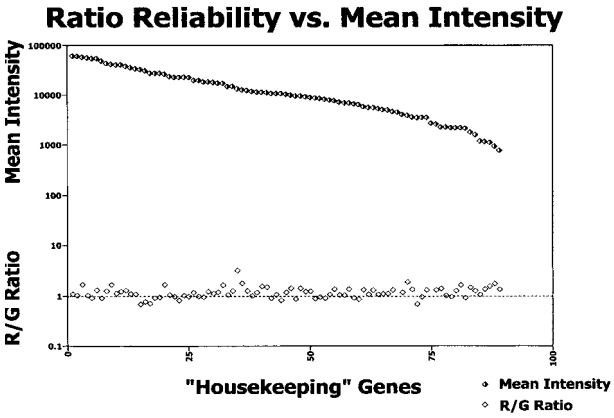

Figure 4. Ratio reliability *vs.* mean intensity graph substantiating the consistent red/green ratios of the "housekeeping" genes (used to compute a normalization factor) over a wide spread of signal intensities

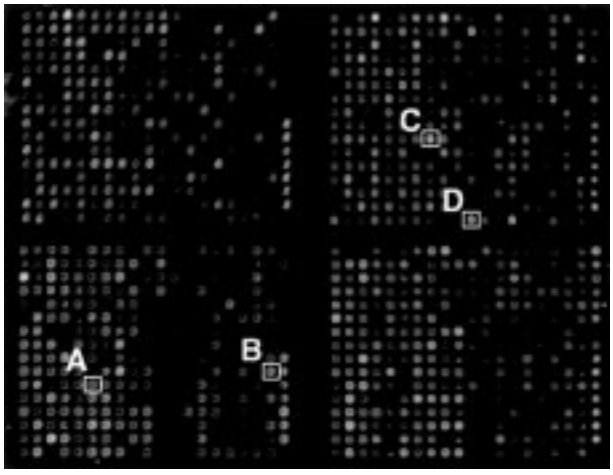

Figure 5. Representative microarray hybridization pseudocolored image generated by the DeArray software. The reference target, green, is a myofibroblast cell line and tumor (rhabdomyosarcoma) is red. The up (red) and down (green) regulation, as well as equal (yellow) expression of several genes are illustrated. Representative genes of interest are boxed: A, *SM22* (R/G ratio of 0.06); B, *ATF3* (39); C, *CAPDH* (0.8); D, *MYCN* (39).

array experiments with northern hybridizations have confirmed this technique to be reliable (see Fig. 6). In the experiment by DeRisi *et al.* [13], mRNA from a tumorigenic melanoma (UACC-903), labeled red, was hybridized against a nontumorigenic derivative of the original cell line (UACC-903+6), labeled green. The red spot represents overexpression of the *MCAF/MCP-1* gene with a ratio > 10, the yellow spot for β-actin gene has an expression ratio of 1, and the green spot for α_1-antichymotrypsin has a ratio of 0.1. The adjacent northern analysis confirms these changes. Our experience to date has indicated the high reliability of microarray data for determining ratio changes; however, there is some

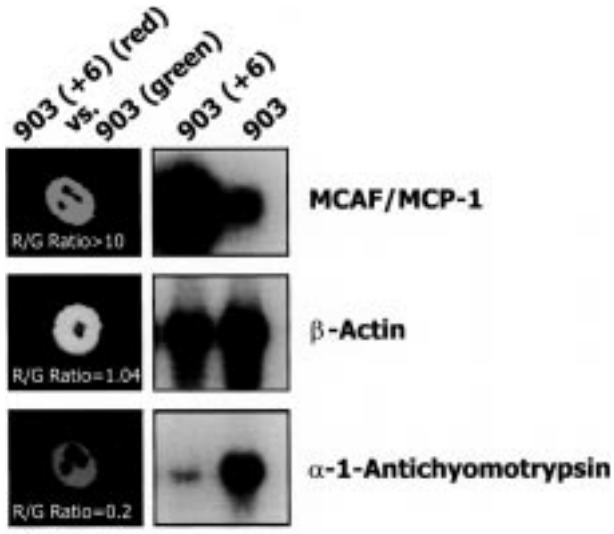

Figure 6. Northern hybridization substantiating the consistency of the cDNA microarray expression ratios. The signal detected by a radio-labeled β-actin probe was used as a control for loading variance, with a red/green ratio observed on the cDNA microarray for β-actin of 1.04. Images were used with permission.

variation in the exact value of the ratios obtained by these two methods. In some instances the ratio obtained by microarray analysis underestimates that obtained by northern analysis. Possible causes for this underestimation include reaching a target intensity saturation limit at the highest intensities under our current detection system. Additionally, the largest ratio changes frequently have one of the measurements near the lower limit of detection, and at these levels some imprecision may occur, causing variance at the higher ratio measurements. Other causes of discrepancy may be due to nonlinear binding characteristics of target to probe. Further investigations, including absolute signal intensity response curves and evaluation of the concordance of microarray measurements with other methods of ratio determination, are important ongoing efforts in the development of this technique.

4 Discussion

More than 1 000 000 EST sequences are available in public databases representing perhaps 40 000–50 000 of the estimated 80 000–100 000 human genes. These data, in conjunction with full-length cDNA sequencing and genomic sequences, will lead to a full catalog of all genes. Although not yet feasible, improvements in current technology should enable the generation of microarrays containing the entire human gene complement. Thus, the expression level of every human gene could be monitored in a single experiment. Already, using currently available cDNA microarrays, the two-color hybridization technique

has been successfully applied to build expression profiles of complex disease processes and metabolic pathways. Heller *et al.* [16] used cDNA microarrays to compare between tissue samples of rheumatoid arthritis and inflammatory bowel disease and identified novel involvement of genes, including cytokine interleukin-3, chemokine Gro alpha and metalloproteinase matrix metalloelastase, in these diseases.

DeRisi *et al.* [17] have used cDNA microarrays containing almost every gene of *Saccharomyces cerevisiae* to decipher the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The same microarrays were used to identify genes whose expression was affected by deletion of the transcriptional corepressor TUP1 or overexpression of the transcriptional activator YAP1. They confirmed the utility of this approach to investigate the metabolic and genetic control of gene expression on a genome-wide scale. This technique has also been used in the identification of known and novel heat shock and phorbol ester-regulated genes in human T cells [18].

The value of microarray technology is not just that that it allows screening for the expression of individual genes, but that it enables the study of global gene-gene interactions. Thus it is possible to examine the status of entire pathways as they are impacted by various manipulations. For example, by introducing a gene into model systems it will be possible to study the downstream effects of transcription factors, oncogenes, and tumor suppressor genes. Microarray analysis will be an invaluable tool for deciphering the complex network of interactions of genes involved in cell cycle, signal transduction, and apoptosis. In addition, a large proportion of genes are represented by anonymous ESTs with no known function; therefore, changes in expression levels of these unidentified genes may lead to further insight into their function.

Studying global gene expression of different types of cancer may allow the development of expression profiles unique for a cancer [19] and may lead to the development of rapid diagnostic assays. It may also identify secreted proteins that can be used for early diagnosis and for monitoring therapy. Gene expression profiles can also be correlated with clinical data to help predict biological behavior, and may allow us to direct therapy. In addition, this information may be useful in dissecting out the pathways involved in malignant transformation and may ultimately provide novel therapeutic targets. Interest in microarray technology has risen in the pharmaceutical industry for new cancer drug discovery and for monitoring the effects of novel therapeutic agents. The list of potential uses of this technique is not limited to cancer research. We envisage that cDNA microarrays will have a major impact on biomedical research that will greatly increase our understanding of all aspects of human disease.

The custom-built robotic arrayer was developed by Stephen B. Leighton and scanner optics by Paul D. Smith. Thomas Pohida developed the electronics of both the arrayer and scanner. We thank Yuan Jiang, Gerald C. Gooden, John Lueders, Kim A. Gayton, Art A. Glatfelter and Robert L. Walker for their excellent technical assistance.

Received November 12, 1998

5 References

[1] Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, D., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, C., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morisette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wlilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehjia, A., Goodfellow, P. N., Houlgatte, R., Hudson Jr., J. R., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Philips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., Deloukas, P., Lander, E. S., Hudson, T. J., *Science* 1996, *274*, 540–546.

[2] Guyer, M. S., Collins, F. S., *Proc. Natl. Acad. Sci. USA* 1995, *92*, 10841–10848.

[3] Rowen, L., Mahairas, G., Hood, L., *Science* 1997, *278*, 605–607.

[4] Vogelstein, B., Fearon, E. R., Hamilton, S. R., Ker, S. E., Preisinger, A. C., Leppert, M., Nakamura, Y., White, R., Smits, A. M., Bos, J. L., *N. Engl. J. Med.* 1988, *319*, 525–532.

[5] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, *270*, 484–487.

[6] Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W., Vogelstein, B., *Nature* 1997, *389*, 300–305.

[7] Drmanac, S., Stavropoulus, N. A., Labat, I., Vonau, J., Hauser, B., Soares, M. B., Drmanac, R., *Genomics* 1996, *37*, 29–40.

[8] Drmanac, S., Drmanac, R., *BioTechniques* 1994, *17*, 328–329, 332–336.

[9] Schena, M., Shalon, D., Davis, R. W., Brown, P. O., *Science* 1995, *270*, 467–470.

[10] Lockhart, D. J., Dong, H., Byrne, M. C., Folletie, M. T., Gallo, M. G., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, L. B. *Nature Biotechnol.* 1996, *14*, 1675–1680.

[11] Boguski, M. S., Schuler, G. D., *Nature Genet.* 1995, *10*, 369–371.

[12] Shalon, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, *6*, 639–645.

[13] DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., Trent, J. M., *Nature Genet.* 1996, *14*, 457–460.

[14] Chen, Y., Dougherty, E. R., Bittner, M. L., *Biomed. Optics* 1997, *2*, 364–374.

[15] Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M., Boguski, M. S., *Nature Genet.* 1998, *20*, 19–23..

[16] Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. E., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, *94*, 2150–2155.

[17] DeRisi, J. L., Iyer, V. R., Brown, P. O., *Science* 1997, *278*, 680–686.

[18] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1996, *93*, 10614–10619.

[19] Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S. B., Pohida, T., Smith, P. D., Jiang, Y., Gooden, G. C., Trent, J. M., Meltzer, P. S., *Cancer Res.* 1998, *58*, 5009–5013.